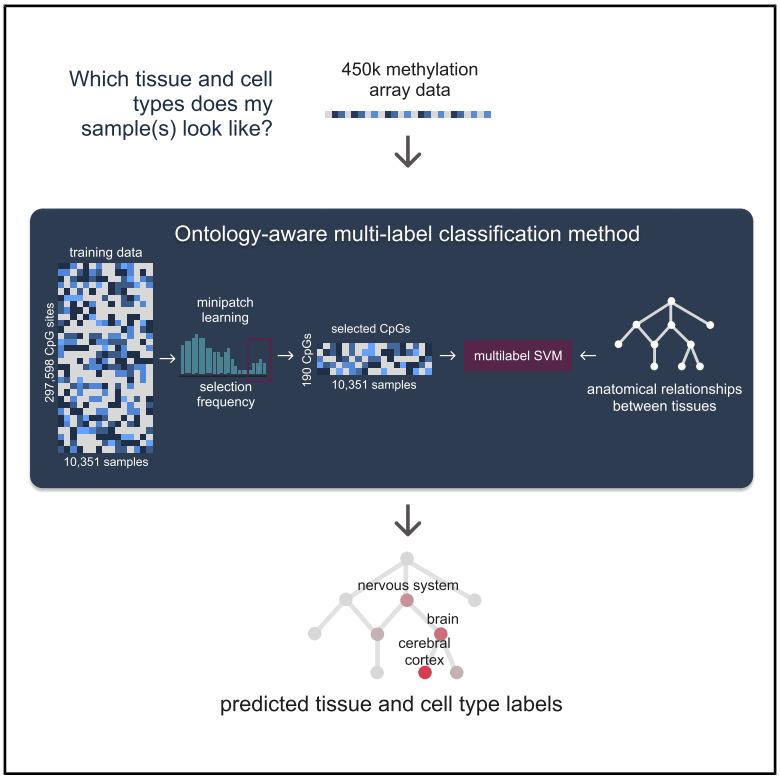
Cell Reports
Methods

Article

Ontology-aware DNA methylation classification with a curated atlas of human tissues and cell types

Graphical abstract



Authors

Mirae Kim, Ruth Dannenfelser, Yufei Cui, Genevera Allen, Vicky Yao

Correspondence

vy@rice.edu

In brief

Kim et al. assemble a large atlas of healthy human DNA methylation and introduce an ontology-aware model that learns hierarchical tissue identity. By identifying a compact CpG signature of normal tissue identity, the work offers a reference baseline that could complement methylation aging clocks in future biological and clinical applications.

Highlights

- Curated atlas of 17k healthy human methylomes across 86 tissues and cell types
- Ontology-aware model learns hierarchical tissue and cell relationships
- Minipatch learning identifies 190 CpGs for accurate multilabel classification
- Model generalizes to unseen biological labels using ontology proximity

Kim et al., 2026, Cell Reports Methods 6, 101328
March 23, 2026
https://doi.org/10.1016/j.crmeth.2026.101328

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

Ontology-aware DNA methylation classification with a curated atlas of human tissues and cell types

Mirae Kim,[1] Ruth Dannenfelser,[1] Yufei Cui,[2] Genevera Allen,[3,4,5,6] and Vicky Yao[1,7,8,9,*]
[1]Department of Computer Science, Rice University, Houston, TX 77005, USA
[2]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, US
[3]Department of Statistics, Columbia University, New York, NY 10027, USA
[4]Center for Theoretical Neuroscience, Columbia University, New York, NY 10027, USA
[5]Zuckerman Institute for Mind, Brain and Behavior, Columbia University, New York, NY 10027, USA
[6]Irving Institute for Cancer Dynamics, Columbia University, New York, NY 10027, USA
[7]Ken Kennedy Institute, Rice University, Houston, TX 77005, USA
[8]Rice Synthetic Biology Institute, Rice University, Houston, TX 77005, USA
[9]Lead contact
*Correspondence: vy@rice.edu
https://doi.org/10.1016/j.crmeth.2026.101328

MOTIVATION A major obstacle in studying healthy tissue- and cell-specific methylation patterns has been the absence of comprehensive, curated reference atlases across diverse normal human tissues. Most existing work has focused on disease-subtype differentiation or localized tissue comparisons, which limits broader biological insight. By assembling a large healthy tissue atlas and developing an ontology-aware classification framework, we aim to enable robust identification of CpG features associated with tissue and cell identity and to improve label transfer for tissues not represented during training. This supports future applications in disease detection and personalized medicine.

SUMMARY

DNA methylation is a key regulatory mechanism reflecting both short- and long-term biological stimuli. While it has been widely used to study aging through disease-associated methylation shifts, its potential for revealing tissue-specific shifts remains underexplored due to the lack of comprehensive reference atlases with correspondingly systematic analysis framework. To address this, we assemble the largest and most diverse atlas of healthy human tissue and cells profiled by 450K arrays, totaling 16,959 samples across 86 tissues and cell types. Using this resource, we introduce an ontology-aware classification framework that identifies robust CpG features linked to tissue and cell identity and incorporates known anatomical and functional relationships. Through minipatch learning, we distill 190 CpGs that support accurate multilabel classification and validate the approach with ontology-based label transfer to 31 unseen tissue and cell types.

INTRODUCTION

DNA methylation is a key epigenetic mechanism responsible for regulating gene expression and chromatin organization, serving both to preserve cell lineage identity and dynamically mediate cellular responses to environmental stimuli.[1] Due to these dual roles, DNA methylation patterns have emerged as powerful biomarkers for a variety of biological processes, including tissue- and cell-type classification based on conserved methylation signatures,[2,3] quantification of aging-associated methylation shifts,[4,5] and detection of molecular alterations associated with disease[6,7] or environmental exposures.[8,9] These broad detection abilities have led to a flurry of excitement about the potential of DNA methylation as a comprehensive molecular snapshot of human health, with tissue- and cell-type classification methods as central components of this vision.[10,11]

Existing tissue and cell classification approaches typically aim to identify stable sets of unique, distinguishing methylation patterns. These patterns are often localized to unmethylated regulatory regions that are important in defining cellular identity and function.[1] Because cell lineage is a primary driver of these conserved methylation signatures, considerable effort has been devoted to computational deconvolution methods that infer cell-type proportions from bulk-tissue methylation

profiles.[10,12] Despite advances in reference-free deconvolution approaches, identification of reliable tissue- and cell-type markers still rely heavily on statistical analyses of DNA methylation reference datasets.[13]

The effectiveness of these methods is closely tied to the comprehensiveness and expressiveness of their underlying reference datasets. Ideally, reference datasets would capture extensive tissue- and cell-type diversity with broad coverage of methylation sites across the genome. In practice, researchers must balance the tradeoff between array-based technologies (e.g., 450K, EPIC arrays), offering broad sample diversity at lower genomic coverage, and sequencing-based technologies (e.g., whole genome bisulfite sequencing [WGBS] and reduced representation bisulfite sequencing [RRBS]), which provide greater genomic coverage across CpG sites but at higher costs and lower sample throughput. Recent advances in single-cell DNA methylation techniques promise unprecedented cellular resolution,[14,15] but their high cost, limited scalability, and intrinsic data sparsity currently limit their use for large-scale tissue profiling.[16,17] Consequently, comprehensive bulk-tissue methylation reference collections remain essential. Various DNA methylation tissue- and cell-type atlases have been assembled across arrays,[18,19] RRBS,[20–22] and WGBS.[2,3] However, some of these efforts combine both healthy and diseased samples to achieve broader tissue coverage, but can inadvertently obscure signals specific to healthy cellular states.[23] Meanwhile, even the most comprehensive healthy atlas, profiled using WGBS, covers only 39 cell types from 18 major tissues,[3] missing entire organ systems (e.g., the male reproductive system) and is limited to a few representative cell types per tissue.

Here, we address these gaps by assembling, to our knowledge, the largest curated atlas of exclusively healthy, primary human tissue and cell types. Our data compendium spans 55 tissue and cell types, sourced from 210 publicly available studies profiled by 450K arrays within the Gene Expression Omnibus (GEO).[24] Although 450K arrays profile fewer CpG sites than WGBS, they still robustly capture cell-type specific methylation signals,[25] and their widespread usage has resulted in unmatched sample diversity. Leveraging this comprehensive atlas, we introduce a multilabel, ontology-aware classification framework explicitly designed to prioritize tissue identity rather than cell-type composition. By focusing on samples from healthy individuals during training, the model learns normal methylation patterns, which provides a clean reference against which disease-associated changes can be studied. Unlike traditional deconvolution methods, our ontology-driven approach incorporates known functional and lineage relationships between tissues and cell types, enabling identification of CpG sites that represent distinct biological signatures beyond lineage origin alone. By anchoring our approach within a structured anatomical and functional ontology, we can also effectively model intermediate nodes and organ-system-level relationships, ultimately expanding our classification coverage to 72 distinct anatomical entities. Our framework leverages ontology both during training, via label propagation, and during interpretation, where predictions are contextualized within the ontology. This dual use enables evaluation of tissues and cell types not seen during training, with accurate predictions for 31 unseen labels, demonstrating robust and biologically meaningful generalizability despite inherent gaps in available reference data.

RESULTS

Curating a diverse DNA methylation compendium of 86 healthy primary tissues and cell types

We assembled a reference compendium of DNA methylation profiles across a wide variety of healthy, untreated, primary tissues, and cell types to model DNA methylation relationships in the context of anatomical and functional similarity. To this end, we obtained all publicly available Illumina 450K data deposited in GEO[24] and manually filtered out samples that were diseased, treated, or derived from cell lines or organoids. We then reprocessed all files using a consistent data processing pipeline, removing samples that did not pass quality control checks (see STAR Methods). For consistency, we disambiguated sample type labels by manually reconciling them with the UBERON Tissue Ontology,[26] yielding a final set of 16,959 samples spanning 86 unique tissue and cell types (Table S1). To enable robust learning, we required that any tissue or cell label be supported by at least two unique studies. This filtering resulted in a smaller set of 10,351 samples across 55 tissues, which, when considering ontological structure (intermediate tissue and organ system level terms), was further expanded to 72 entities (Figure 1). The remaining 6,608 samples covering 31 tissue and cell types were reserved for the label transfer evaluation. This effort represents, to our knowledge, the largest and most diverse DNA methylation atlas of healthy primary tissues and cell types, providing an unprecedented resource for exploration methylation patterns in normal physiology. Such a resource is critical not only for training our ontology-aware classification framework, but also for enabling future studies on tissue-specific epigenetic regulation and its implications for human health.

To assess the consistency of our compendium, we examined *M*-value[27] distributions across sample types and datasets (Figure 1A) and quantified probe variation using intraclass correlation (ICC). In general, samples of the same type exhibited similar methylation patterns (mean ICC = 0.97; Figure S1 and Table S2), regardless of the number of samples per label. Sample types with greater variability likely reflect known tissue heterogeneity or cell types present in diverse bodily regions, such as the two sample types with the lowest ICCs, "bone marrow" (ICC = 0.92; $N = 74$) and "fibroblast" (ICC = 0.93; $N = 117$). Interestingly, some heterogeneous tissues like "breast" and "lung", which both contain different mixtures of diverse cell types (e.g., fibroblasts, epithelial cells, immune cells, as well as adipocytes in the breast), still exhibited high consistency (breast: ICC = 0.98, $N = 314$; lung: ICC = 0.97, $N = 68$), suggesting that robust tissue-specific signals can be extracted without deconvolution. Notably, "neutrophils" (ICC = 0.99; $N = 68$) and "umbilical cord" (ICC = 0.99; $N = 1,021$) had the highest internal consistency across all 450K probes.

Next, we analyzed the relationships between tissue and cell types by extracting known anatomical and functional relationships, using *is_a*, *part_of*, and *develops_from* annotations from the UBERON tissue ontology.[26] We combined automated extraction from UBERON with pruning and simplification of the

CellPress
OPEN ACCESS

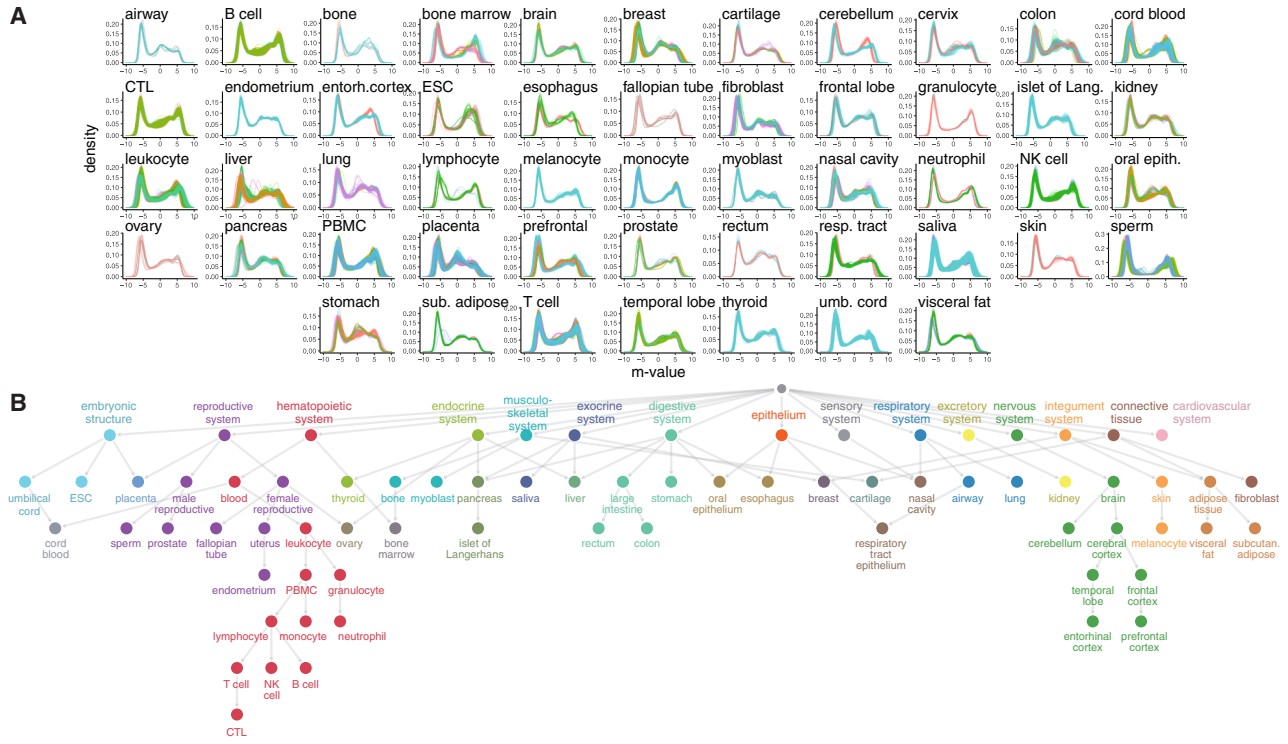


Figure 1. Overview of DNA methylation training compendium

(A) Distribution of processed *M*-values across all probes for all samples used for downstream learning tasks. Each line in the density plot corresponds to an individual sample, with colors indicating different datasets.

(B) Anatomical ontology structure after manual curation and subsetting to tissue and cell types in the training set. Nodes are colored by their organ system, and these colors are used throughout the remainder of the paper.

See also Table S1.

network, as well as other functional *is_a* and *part_of* relationships from the BRENDA tissue ontology.[28] This process yielded a directed acyclic graph spanning all sample types in our compendium, organized with increasing levels of physiological resolution (Figures 1B and S2). Our tissue and cell ontology captures structural and functional heterogeneity through multiple parent nodes. For example, "breast" is a third level node with two organ system-level parents: "exocrine system" and "connective tissue". Structurally, cell types are typically at deeper levels of the ontology, with relationships traceable back to the relevant parent organ system. For instance, the deepest node, "cytotoxic T lymphocyte (CTL)", is functionally linked to "hematopoietic system", "blood", "leukocyte", "PBMC", "lymphocyte", and "T cell" (Figure 1B).

Minipatch-based feature selection of DNA methylation features

To explore the functional conservation of CpGs across tissues and cell types, we devised an ontology-aware, multilabel classification framework (Figure 2). In our approach, we begin by considering the union of all quality-controlled CpG sites (297,598 sites) across the 10,351 training samples in a 3-fold cross-validation setup. Importantly, cross-validation folds were grouped by study, such that samples from the same dataset were always allocated to the same fold, to ameliorate the possibility of data leakage. A key innovation of this framework is the use of minipatch learning,[29] a probabilistic sampling approach for feature selection, which iteratively refines the feature space down to a set of 190 CpG sites (Table S3). Briefly, minipatch learning uses different "patches" within the data by sampling small subsets of CpGs and samples and iteratively evaluates feature importance using decision tree classifiers over many minipatches. The sampling probability of CpG features is increased based on their feature importance (i.e., "exploited"), while the broader feature space continues to be "explored." After iterating, the selection frequency of a feature thus directly reflects its classification relevance. By iteratively identifying CpGs that consistently contribute to accurate classification across many minipatches, the feature selection process emphasizes data-driven and generalizable feature relevance rather than markers specific to a single tissue or cell type.

We optimized the selection frequency cutoff based on the elbow (frequency = 0.65) at which the median F1 scores on the downstream multilabel sample type classification task starts to decrease across each of the cross validation folds (fold 1 F_1 = 0.80; fold 2 F_1 = 0.90; fold 3 F_1 = 0.87; Figure 3A). More stringent selection frequency cutoffs result in substantial performance decreases. This strategy is effective and efficient, enabling sample classification approximately 200× faster compared to the traditional differential methylation approach (Figure 3B), while

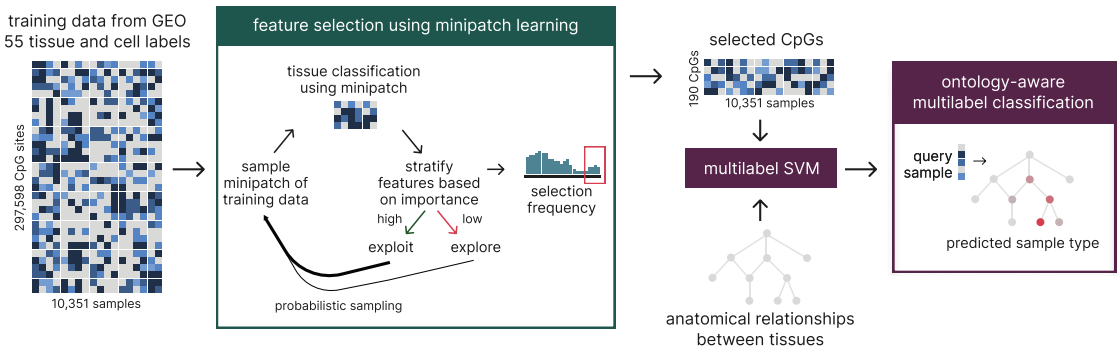


Figure 2. DNA methylation feature selection and classification workflow

We first used minipatch learning[29] for feature selection, reducing the number of CpG features from 297,598 to 190. Selected probes are then combined with the anatomical relationships between tissues and cell types in a multilabel SVM learning framework. Given a sample of unknown origin, our ontology-aware classification framework is capable of assigning the most relevant label.

offering support for multilabel input. We note that the runtime difference compounds over large sample sizes, making it increasingly intractable to use a differential methylation-based classification approach on large data compendia.

To further examine the CpGs selected via minipatch learning, we visualized the DNA methylation feature space using principal component analysis (PCA). We found that the complete feature space of 297,598 CpGs (Figure 3C) provided some separation between sample types, but the PCA based on the 190 minipatch learning-selected probes not only maintained but amplified this separation (Figure 3D). We also applied supervised *k*-means clustering to the PCA space to assess whether tissue labels could be reliably separated, but the resulting median tissue-wise precision was low (0.101), indicating that unsupervised clustering alone is not sufficient for this task (Tables S5 and S6). Furthermore, this reduced feature set captured substantially more variance in the first two principal components, highlighting that these probes effectively capture sample-type relevant signal.

We also characterized the genomic composition of the 190 selected CpG sites, which span all chromosomes except for chromosome 18 (Figure 3E; Table S3). Interestingly, although the collection of probes in 450K arrays are biased toward CpG-rich regions such as CpG islands, these regions are notably underrepresented in the selected set (one-sided Fisher's exact test, $p = 1.9*10^{-13}$). Conversely, there is a significant over-enrichment of CpGs in open sea (one-sided Fisher's exact test, $p = 1.2*10^{-6}$) and shelf regions (one-sided Fisher's exact test, $p = 2.0*10^{-3}$), suggesting that these areas may be more useful for tissue- and cell-type specificity. We observed no significant enrichment or depletion in the shore regions (two-sided Fisher's exact test, $p = 0.6$). This is consistent with previous reports analyzing tissue-specific differentially methylated regions, finding CpGs to be predominantly located outside of islands[30,31] and specifically more localized to shelves and also distant regions in 450K data,[18] where changes are tied with alternative transcription.[32] To functionally characterize the selected CpGs, we mapped each CpG to nearby genes and regulatory regions, and performed gene set enrichment analysis. The results show enrichment for core regulatory and transcriptional processes (e.g., "positive regulation of DNA-templated transcription," "sequence-specific DNA binding"; adjusted $p < 0.05$; Table S4), suggesting that the selected CpGs capture broadly relevant regulatory features that may mark key elements defining and distinguishing tissue identities. We also note that the relative distribution of selected CpGs per chromosome deviates from that expected by CpG site abundance alone, indicating that the selection is not simply driven by CpG site availability or chromosome length.

Accurate cell, tissue, and organ system classification with our ontology-aware framework

To integrate ontology awareness into the framework, we used a label propagation strategy where each sample's label includes its annotated specific label as well as all of its parent terms (Figure 1). For example, for the "leukocyte" class, the propagated label set includes "leukocyte", "blood", and "hematopoietic system". This propagated label set is used directly as input for training a multilabel support vector machine (SVM) classifier, enabling predictions across the full ontology and capturing both broad organ systems and specific tissue or cell types, making the training process itself ontology-aware. During prediction, each label is assigned a probability using Platt scaling, and labels with probabilities above 0.50 are considered positive. The resulting set of positive labels is then mapped back onto the ontology, enabling interpretation across hierarchical levels. If no label exceeds the threshold, the model returns "no prediction," reducing false positives in low-confidence cases.

Using 3-fold cross-validation, we applied this ontology-aware classification framework to all 10,351 samples and compared its performance to a naive correlation-based baseline that uses tissue- and cell-type-specific probes derived from differential methylation (Figure 4A; Tables S5 and S6). In this baseline, significantly differentially methylated CpGs (Holm-Sidak, $\alpha = 0.05$) unique to each tissue are combined into a union set, which is then used to calculate sample-to-sample correlations, and the correlation values are used as proxies for prediction probabilities. To allow a lenient comparison, both methods were evaluated against the propagated label sets of the reference annotations, such that predictions of parent terms were considered correct. Across the 51 training labels that had sufficient annotations for baseline correlation analyses, our method had

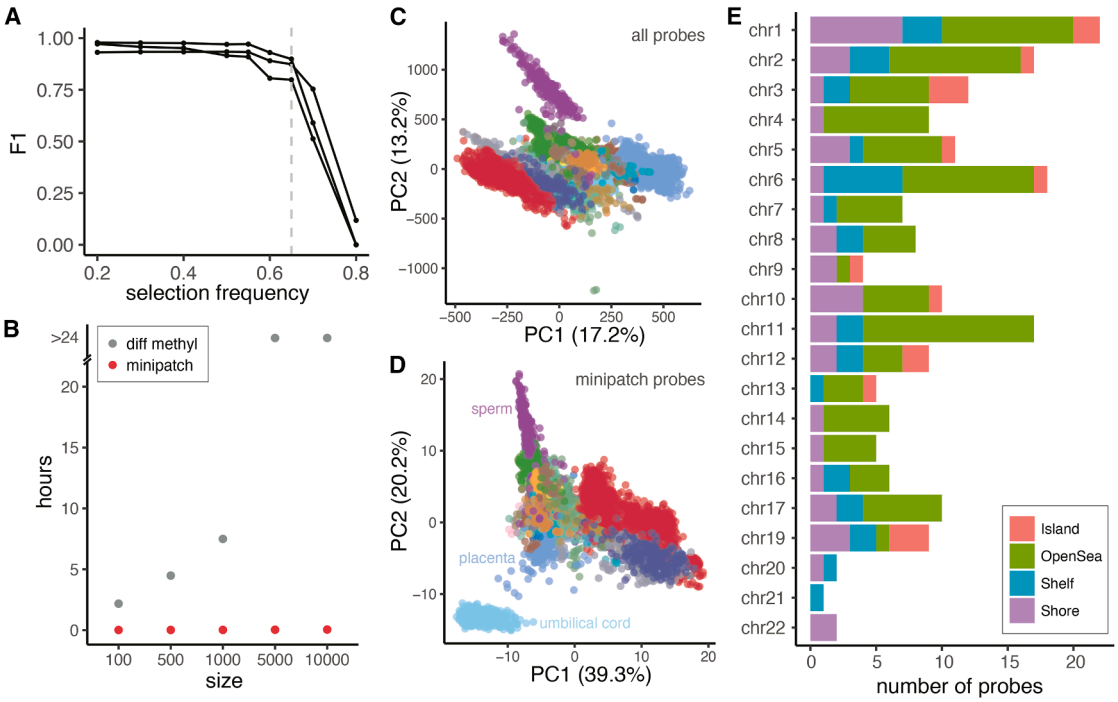


Figure 3. Characterization of selected CpGs for downstream classification

(A) Overall classification performance (F1 score) as a function of CpG selection frequency. Each line represents one training fold. A selection frequency of 0.65 (dotted line) was chosen to maximize performance while minimizing the total number of CpGs.

(B) The runtime, in hours, as a function of the sample size of the tissue- and cell-type classification task when using minipatch probes versus the set of differentially methylated probes. For sample sizes of 5,000 and above, the runtime for differential methylation was greater than 24 h.

(C and D) PCA of samples using the entire probe set (C) and the reduced set after minipatch learning (D). Colors correspond to the organ system as in Figure 1B.

(E) Chromosome and genomic features characterizing the 190 minipatch learning-selected probes.

See also Table S3.

significantly higher area under the precision-recall curve (AUPRC) (one-sided Wilcoxon signed-rank test, $p = 4.13*10^{-4}$). Furthermore, our ontology-aware model was able to leverage the hierarchical relationships when direct annotations were unavailable to make predictions for an additional 21 tissues and cell-type terms (median AUPRC = 0.927). Even in tissues with lower AUPRC, such as "neutrophil", "temporal lobe", and "prefrontal cortex", the most frequent alternative predictions corresponded to biologically related parent or sibling terms (e.g., "hematopoietic system", "blood", "leukocyte", and "granulocyte" for "neutrophil"), indicating that apparent misclassifications still reflected meaningful relationships within the ontology.

To isolate the sources of performance gain and compare against conventional machine learning methods, we performed ablation analyses targeting both the feature selection and classification components of our framework. Each model was evaluated under three complementary evaluation settings to ensure fair comparison across architectures and to contextualize performance in the ontology: (1) propagated, where parent predictions were counted as correct; (2) propagated-neutral, where parent predictions were ignored (not considered correct or incorrect); and (3) single-label, where only the directly annotated tissue was considered correct and parent predictions were treated as incorrect. As expected, our ontology-aware model performed best in the propagated and propagated-neutral settings, where predictions consistent with the ontology are rewarded, and lower in the strict single-label setting, where predicting-related tissues is penalized (Tables S5 and S6). In this single-label evaluation, both multilabel SVM variants (minipatch learning and differential methylation feature sets) achieved slightly lower AUPRC than the correlation-based baseline, reflecting the model's tendency to assign higher probabilities to biologically related parent or sibling labels rather than solely the exact annotation. Meanwhile, conventional multiclass machine learning classifiers (elastic net, random forest, or MLP) provided clear performance improvements over correlation-based methods in this setting. Once hierarchical relationships were incorporated, however, the multilabel SVM with minipatch learning feature selection achieved the highest overall AUPRC and precision across models. Minipatch learning as a feature selection strategy consistently improved tissue-wise AUPRC and precision compared to models trained on top differentially methylated probes across tissues. We note that the ablation analyses were limited to the 51 training labels with adequate annotations, while the ontology-aware approaches also generated high-quality predictions (median AUPRC = 0.927) for 21 tissues and cell types lacking sufficient direct annotations for the ablated methods.

To examine how sample size influences classification performance, we evaluated the AUPRC for each tissue label as a function of the number of samples (Figure 4B). Though many tissues

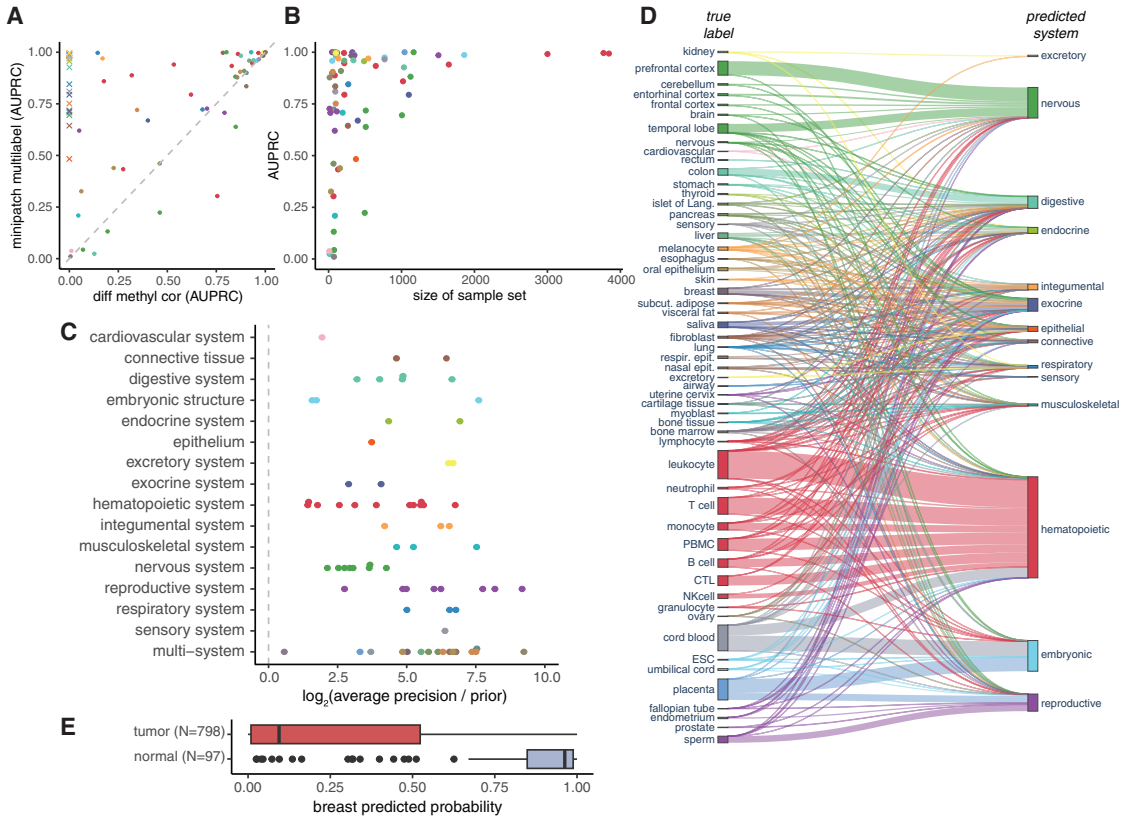


Figure 4. Multilabel SVM performance across tissue and cell types

(A) Tissue-wise area under the precision-recall curve (AUPRC) compared to a baseline method that assigns labels based on correlation with samples using the union of probes that were significantly differentially methylated. Each dot represents the average AUPRC for a single tissue or cell type across folds, colored by its organ system. Tissues drawn as "x"s are intermediate nodes that cannot be predicted using differential methylation due to the lack of directly annotated labels.

(B) Tissue-wise AUPRC as a function of the number of samples with the corresponding tissue label in the training set. Colors correspond to organ system as in Figure 1B.

(C) Log average precision over prior for each tissue- and cell-type label, organized by organ system. The dotted line indicates performance equal to prior.

(D) Sankey plot of the actual sample label compared with the predicted system label colored according to the true organ system. The size of the Sankey ribbon corresponds to the sample size of each label.

(E) Boxplot of predicted probabilities for the "breast" label on TCGA breast cancer patient samples, separated by tumor versus adjacent normal breast tissue, shows significantly lower probabilities in tumor samples (Mann-Whitney U = 68118.0; p value = 2.016e-34).

See also Tables S5 and S6.

and cell types achieved strong performance with relatively few samples, performance generally increased and stabilized at larger sample sizes, especially once the number of samples per label exceeded 1,000 (median AUPRC = 0.985). All but one of the tissue or cell types with lower than 0.5 AUPRC had fewer than 400 samples. When accounting for class imbalance by comparing performance against label priors (the proportion of positive labels), all 72 tissues in the ontology outperformed their respective priors (Figure 4C). We further investigated how the number and organization of candidate labels influenced classification performance using two data subsetting strategies (Figure S4). Under the first strategy, stratified subsets of our data compendium maintained the same distribution of organ systems while varying the proportion of samples. We found that performance improved with increasing subset size, suggesting that shared methylation features among related tissues within each system helps facilitate learning and generalization. In the second strategy, system-based subsets varied the number of organ systems and therefore the number of candidate labels. Performance remained mostly stable, but with a slight decline as the number of systems increased, reflecting the added complexity of learning across a broader label space.

To assess our model's ability to capture broader ontological relationships, we analyzed predictions at the organ system level by comparing the predicted system nodes to true labels. Visualizing this on a Sankey diagram connecting true and predicted labels, colored according to their true systems, we see that the overwhelming majority of predictions are color-coherent, indicating that predictions either match the true label or belong to the same system (Figure 4D). This analysis also clearly highlighted multisystem nodes, where a single tissue or cell label is associated with two or more organ systems, as shown by their mixed colored ribbons in the Sankey diagram. Notable examples include "breast", "nasal epithelium", "placenta", and "umbilical

cord blood''; for instance, ''umbilical cord blood'' correctly maps to both the ''hematopoietic'' and ''embryonic'' systems, while ''placenta'' is also linked accurately to both the ''embryonic'' and ''reproductive'' systems.

Because disease development is often accompanied by methylation shifts away from healthy tissue patterns, we next evaluated whether the classifier captures these deviations using samples from TCGA breast cancer patients,[33] including both primary tumor tissue and histologically normal adjacent breast tissue (Figure 4E). We observe that ''breast'' prediction probabilities for normal samples remain consistently high, whereas those for tumor samples are significantly lower (Mann-Whitney U = 68118.0; p value = 2.016e-34), suggesting that the model is sensitive to normal tissue methylation profiles and can capture disease-associated drift from healthy states.

We also evaluated whether the framework remains robust across biological and technical differences on independent datasets not used during training. This included blood samples with different cell compositions[34] and pancreatic samples profiled using different array technologies.[35] The blood dataset consists of artificially mixed leukocyte fractions and purified immune cell types, providing samples with known proportions of granulocytes, lymphocytes, and monocytes.[34] Regardless of cell type composition, all samples were confidently predicted as ''blood''. Interestingly, the prediction probabilities to some extent reflected underlying cell-type composition. ''Granulocyte'' fractions showed strong correlation with corresponding predicted probabilities (Spearman = 0.95, p value = 2.92e-9), $CD4^+$ and $CD8^+$ T cell fractions correlated with ''T cell'' prediction probabilities (Spearman = 0.79, p value = 1.1e-4), and ''natural killer cell'' fractions also correlated significantly (Spearman = 0.70, p value = 0.0012; Table S7). When examining the independent pancreas dataset[35] that included samples profiled on both 450K and EPIC (850K) arrays, our model returned confident ''pancreas'' predictions for all samples, with comparable prediction probabilities across platforms (Table S7). Though our framework was trained using only samples profiled using 450K arrays, this analysis suggests that the learned representations are stable across array technologies with differing CpG coverage.

Ontology-aware learning enables robust predictions on unseen labels

A major advantage of our method is the ability to leverage the ontology to predict relevant, related sample type labels, even when the original sample type is absent from the training data. This capability is valuable in real-world contexts where metadata labels may be incomplete, inconsistent, or contain errors, and in large-scale or clinical studies where there may be rare or novel tissues. To evaluate this capability, we used our ontology-aware classification model to make predictions for all samples that were curated as part of our data resource but were excluded from training due to having insufficient numbers of samples or distinct studies. This resulted in a set of predictions for 6,608 samples across 31 unseen tissue and cell type labels, spanning 11 organ systems (Table S1). Because the true target labels were not part of our training ontology, we devised an evaluation metric that measures how close each prediction is to the true target label if it were incorporated into the training ontology (Figures 5A and S2). Specifically, we used an adjusted ontology distance metric that calculates each sample's average graph distance from its predicted labels to the true label, accounting for the minimum distance from any training node to the target label. An optimal prediction would have a distance of 0, with larger distances indicating more disparate predictions. To account for variations in the ontology structure, we sampled 1,000 random nodes from our training ontology and measured their distances to the target node as a baseline. Because these unseen labels have no annotated examples in the training data, standard classifiers or transfer learning approaches are not directly applicable. The random label baseline therefore provides a fair performance reference, directly quantifying whether our ontology-aware framework captures predictive signals beyond chance.

Our method significantly outperforms the random baseline for 22 out of the 31 tissue- and cell-type labels (Figure S3). When grouped by organ system, our approach consistently outperforms random predictions in nearly every system, with the exceptions of the ''endocrine'' and ''cardiovascular'' systems (Figure 5C). We speculate that the lower performance in the ''endocrine system'' may stem from its limited representation in the training set, since all endocrine tissues in our dataset are classified under multiple systems. Similarly, the ''cardiovascular system'' likely suffers from both having the least amount of training data overall and the fact that it contains only two multisystem tissue types in its unseen label evaluation set (''endothelial cell'' and ''aorta smooth muscle tissue''). We also observed that, perhaps unsurprisingly, unseen tissues closer in ontology distance to the nearest observed tissue resulted in predicted label sets that were more closely aligned with their true labels (Figure 5B).

To illustrate how our ontology-driven approach can be used to interpret unseen samples, we can consider an individual sample from the ''epithelium of trachea'' (Figure 6, top), a tissue label unseen in our training set. Our method was able to both correctly identify its closest match, ''respiratory tract epithelium'' (probability = 0.87), and capture all relevant organ systems with high probability: ''epithelium'' (probability = 0.99), ''respiratory system'' (probability = 0.99), and ''sensory system'' (probability = 0.99), highlighting the benefit of this multisystem aware prediction framework. We can also use the same prediction scheme to consider the collection of all samples from another unseen label, ''macrophage'' (Figure 6, bottom). Biologically, we know that ''macrophage'' should be a child node of ''leukocyte''. Our classifier not only captured this chain of parent relationships up through ''hematopoietic system'', but also reflected downstream functional links between monocytes and macrophages, as monocytes can differentiate into macrophages when recruited from the blood into tissues.[36] This example also highlights another use case of our method to summarize predictions across a set of samples. In general, our framework's robust ability to place unseen labels in the context of our training data provides additional validation for the benefit of integrating structured ontological knowledge into the classification process.

DISCUSSION

In this study, we developed an ontology-aware multilabel classification framework leveraging DNA methylation data to

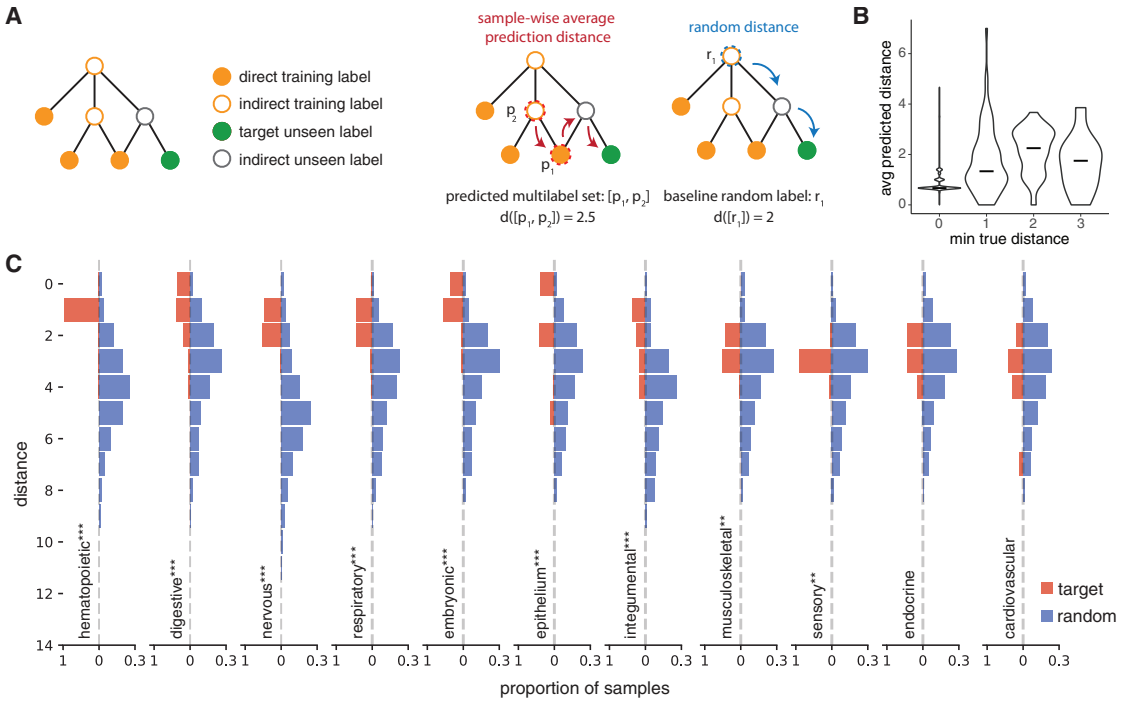


Figure 5. Model performance capturing-related tissues and cell types for samples from unseen labels

(A) Schematic of the label transfer evaluation, given the full tissue and cell ontology, including both label transfer and training set labels (Figure S2). Given a query sample, we calculated the graph distance between every tissue label in the predicted label set (dashed red) and the target, previously unseen label (filled green circle). The final score is an average over all predicted labels for a given sample. In the random distance case, we calculated the distances for 1,000 randomly selected labels to the target label to obtain a background distribution of graph distances.

(B) Violin plots of average adjusted prediction ontology distances per sample, grouped by minimum true ontology distance. For each sample, the adjusted distance is computed as the average ontology distance between all predicted labels and the true unseen label, after subtracting the minimum possible distance. Lower values correspond to more accurate predictions. Minimum distance of 0 indicates specific samples with labels in training ontology but excluded due to study size or blood-annotated samples with no specific cell type.

(C) Histograms show the results of the label transfer evaluation, with samples grouped by organ system and ordered by significance. Sample-wise average prediction distances to the target sample are shown in red, while the background distributions of 1,000 random labels are shown in blue. All distances are adjusted such that the optimal distance when predicted correctly is equal to 0. Higher distances indicate a worse set of multilabel predictions. Asterisks represent significance compared to random using the Wilcoxon rank-sum test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

accurately predict and characterize tissue and cell type identities. By assembling the largest curated atlas to date of exclusively healthy, primary human tissues profiled by 450K arrays, we now provide a valuable resource enabling detailed analyses of epigenetic landscapes across a diverse array of physiological systems. Importantly, our approach identifies a small subset of 190 CpGs sites that robustly distinguishes 72 tissues and cell types, demonstrating that a small number of well-selected markers can achieve high classification performance. These markers thus represent valuable reference points for establishing tissue-specific DNA methylation baselines, potentially aiding in the future interpretation of methylation changes associated with disease or environmental influences.

In addition, while the fact that DNA methylation captures lineage relationships between tissues and cell types has been well-documented,[1] our findings underscore the capacity to extract and leverage functional system-related relationships in DNA methylation data as well. This is evidenced by the high ICC consistency across samples within heterogeneous tissues (Figure S1; Table S2) and the ability of our ontology-guided approach to capture both lineage and functional information, including mechanistically related tissue and cell types in the label transfer evaluation (Figures 5, 6, and S3). Our analysis also revealed that predictive tissue- and cell-specific CpGs are predominantly localized in open sea and shelf regions rather than in CpG Islands, which corroborates previous studies of tissue- and cell-type-specific differential methylation.[18,30,31]

Our integration of structured ontological information enabled the multilabel classifier to incorporate tissue- and cell-type similarity beyond explicit annotations. The clustering observed in PCA reflects the strong tissue-specific organization of DNA methylation, consistent with known biological patterns. However, these clusters are not completely distinct, particularly among related tissues, highlighting the need for approaches that provide quantitative, sample-level predictions. Our classification framework builds on this inherent structure to resolve overlaps and align predictions with the defined ontology, producing interpretable outputs that capture both broad and specific tissue relationships. An interesting extension of our framework could be

CellPress
OPEN ACCESS

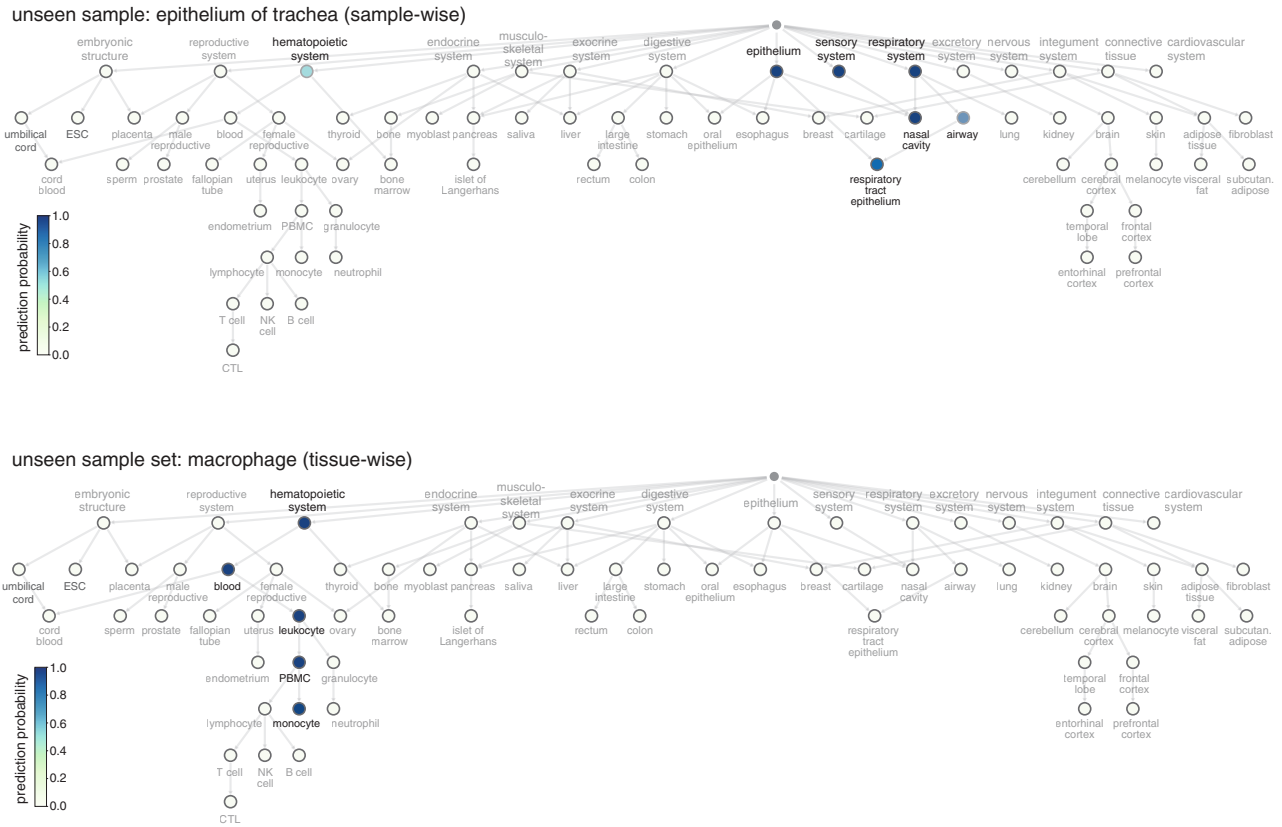


Figure 6. Label transfer predictions for samples with unseen labels

Example multilabel predictions are visualized on the training ontology for two different unseen cases: a single sample with the actual label of "epithelium of trachea" (top) and a set of samples labeled "macrophage" (bottom). Nodes are colored by prediction probability.

to incorporate unsupervised clustering to refine ontology relationships or reveal intermediate and underrepresented tissue states, providing a data-driven avenue to propose updates to existing ontologies.

Our multilabel, ontology-aware setup also allows the model to make predictions at multiple levels of the ontology and to infer both direct and related labels. This was recapitulated with high performance in hierarchy-aware metrics compared to those of both baseline correlation-based methods and conventional machine learning classifiers. Beyond strong classification performance, this approach also offers practical value to biomedical researchers by enabling more accurate annotations, enhancing interpretation of heterogeneously labeled datasets, and providing biologically meaningful predictions for rare or previously uncharacterized tissues, even when explicit labels are unavailable.

Looking forward, an exciting extension of our ontology-based classification framework would be integration with complementary epigenetic data modalities or sparse single-cell DNA methylation data, enabling fine-grained, cell-specific functional analyses in both healthy and disease contexts. As the amount of data increases, we also envision adapting this approach to leverage graph-based or network-aware machine learning techniques, allowing even richer incorporation of complex, multifaceted tissue and cell-type relationships into the classification framework. In a manner analogous to how epigenetic clocks established from healthy individuals have provided valuable baselines for biological age estimation and have yielded critical insights into aging and disease susceptibility,[4,5,37] our comprehensive atlas of healthy tissue- and cell-type methylation profiles establishes a foundational reference for future tissue-based analyses. Ultimately, this resource and ontology-informed modeling approach brings a new perspective to analyses of tissue- and cell-type methylation and paves the way toward deeper insights into tissue-specific disease processes and epigenetic regulation.

Limitations of the study

Though we present the largest DNA methylation data compendium of its kind, we find that sample availability and abundance remains a critical limitation, such as for the cardiovascular and endocrine systems. Currently, our ontology is limited to the sample types available from GEO measured using the Illumina Infinium HumanMethylation450K platform. Expanding data collection and curation efforts to other DNA methylation platforms, including emerging single-cell DNA methylation technologies, would help further provide a richer, multiresolution view of the epigenetic landscape.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Vicky Yao (vy@rice.edu).

Materials availability

This study did not generate new materials.

Data and code availability

- The full data compendium (*M*-values from all 16,959 downloaded, preprocessed, and normalized samples) and annotations (Table S1) are publicly available on HuggingFace at https://doi.org/10.57967/hf/7186.
- All analysis code and the curated ontology are publicly available on GitHub at https://github.com/ylaboratory/methylation-classification and Zenodo at https://zenodo.org/records/17861095 under the BSD 3-clause open-source license.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

The authors would like to thank members of the ylaboratory for helpful discussions. This work was supported by the Cancer Prevention & Research Institute of Texas (CPRIT RR190065 to V.Y.) and the National Science Foundation (NSF DBI-2144534 to V.Y. and NSF DMS-1554821 to G.A.). V.Y. is a CPRIT Scholar in Cancer Research.

AUTHOR CONTRIBUTIONS

Conceptualization, V.Y.; methodology, M.K., G.A., and V.Y.; investigation, M.K., Y.C., R.D., and V.Y.; writing – original draft, M.K. and R.D.; writing – review and editing, M.K., R.D., and V.Y.; funding acquisition, G.A. and V.Y.; resources, V.Y.; supervision, V.Y.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Sample downloading and curation
  - DNA methylation data preprocessing and normalization
  - Sample partitioning for training or label transfer validation
  - Genomic annotation of CpG probes
  - Tissue and cell ontology
  - CpG feature selection using minipatch learning
  - Runtime evaluations
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Within tissue or cell type sample consistency evaluation
  - Hyperparameter optimization and cross-validation
  - Gene enrichment of selected probes
  - Ontology-aware multi-label classification
  - Differential methylation baseline
  - Ablation studies
  - Label transfer evaluation

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2026.101328.

Received: May 13, 2025
Revised: November 4, 2025
Accepted: January 21, 2026
Published: March 12, 2026

REFERENCES

1. Dor, Y., and Cedar, H. (2018). Principles of dna methylation and their implications for biology and medicine. Lancet *392*, 777–786.
2. Ziller, M.J., Gu, H., Müller, F., Donaghey, J., Tsai, L.T.Y., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D.A., Bernstein, B.E., et al. (2013). Charting a dynamic dna methylation landscape of the human genome. Nature *500*, 477–481.
3. Loyfer, N., Magenheim, J., Peretz, A., Cann, G., Bredno, J., Klochendler, A., Fox-Fisher, I., Shabi-Porat, S., Hecht, M., Pelet, T., et al. (2023). A dna methylation atlas of normal human cell types. Nature *613*, 355–364.
4. Horvath, S. (2013). DNA methylation age of human tissues and cell types. Genome Biol. *14*, R115. https://doi.org/10.1186/gb-2013-14-10-r115.
5. Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.B., Gao, Y., et al. (2013). Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. Mol. Cell *49*, 359–367. https://doi.org/10.1016/j.molcel.2012.10.016.
6. Sokolov, A.V., and Schiöth, H.B. (2024). Decoding depression: a comprehensive multi-cohort exploration of blood dna methylation using machine learning and deep learning approaches. Transl. Psychiatry *14*, 287.
7. Yang, Z., Wong, A., Kuh, D., Paul, D.S., Rakyan, V.K., Leslie, R.D., Zheng, S.C., Widschwendter, M., Beck, S., and Teschendorff, A.E. (2016). Correlation of an epigenetic mitotic clock with cancer risk. Genome Biol. *17*, 205–218.
8. Zheng, S.C., Breeze, C.E., Beck, S., and Teschendorff, A.E. (2018). Identification of differentially methylated cell types in epigenome-wide association studies. Nat. Methods *15*, 1059–1066.
9. Teschendorff, A.E., Breeze, C.E., Zheng, S.C., and Beck, S. (2017). A comparison of reference-based algorithms for correcting cell-type heterogeneity in epigenome-wide association studies. BMC Bioinf. *18*, 105–114.
10. Titus, A.J., Gallimore, R.M., Salas, L.A., and Christensen, B.C. (2017). Cell-type deconvolution from dna methylation: a review of recent applications. Hum. Mol. Genet. *26*, R216–R224.
11. Yousefi, P.D., Suderman, M., Langdon, R., Whitehurst, O., Davey Smith, G., and Relton, C.L. (2022). Dna methylation-based predictors of health: applications and statistical considerations. Nat. Rev. Genet. *23*, 369–383.
12. Ji, H., Ehrlich, L.I.R., Seita, J., Murakami, P., Doi, A., Lindau, P., Lee, H., Aryee, M.J., Irizarry, R.A., Kim, K., et al. (2010). Comprehensive methylome map of lineage commitment from haematopoietic progenitors. Nature *467*, 338–342. https://doi.org/10.1038/nature09367.
13. Teschendorff, A.E., and Relton, C.L. (2018). Statistical and integrative system-level analysis of dna methylation data. Nat. Rev. Genet. *19*, 129–147.
14. Luo, C., Keown, C.L., Kurihara, L., Zhou, J., He, Y., Li, J., Castanon, R., Lucero, J., Nery, J.R., Sandoval, J.P., et al. (2017). Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science *357*, 600–604. https://doi.org/10.1126/science.aan3351.
15. Mulqueen, R.M., Pokholok, D., Norberg, S.J., Torkenczy, K.A., Fields, A.J., Sun, D., Sinnamon, J.R., Shendure, J., Trapnell, C., O'Roak, B.J., et al. (2018). Highly scalable generation of DNA methylation profiles in single cells. Nat. Biotechnol. *36*, 428–431. https://doi.org/10.1038/nbt.4112.
16. Iqbal, W., and Zhou, W. (2023). Computational Methods for Single-Cell DNA Methylome Analysis. Genom. Proteom. Bioinform. *21*, 48–66. https://doi.org/10.1016/j.gpb.2022.05.007.
17. Ahn, J., Heo, S., Lee, J., and Bang, D. (2021). Introduction to Single-Cell DNA Methylation Profiling Methods. Biomolecules *11*, 1013. https://doi.org/10.3390/biom11071013.
18. Lokk, K., Modhukur, V., Rajashekar, B., Märtens, K., Mägi, R., Kolde, R., Koltšina, M., Nilsson, T.K., Vilo, J., Salumets, A., and Tõnisson, N.

(2014). Dna methylome profiling of human tissues identifies global and tissue-specific methylation patterns. Genome Biol. *15*, r54.

19. Teschendorff, A.E., Zhu, T., Breeze, C.E., and Beck, S. (2020). Episcore: cell type deconvolution of bulk tissue dna methylomes from single-cell rna-seq data. Genome Biol. *21*, 221–233.
20. Moss, J., Magenheim, J., Neiman, D., Zemmour, H., Loyfer, N., Korach, A., Samet, Y., Maoz, M., Druid, H., Arner, P., et al. (2018). Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free dna in health and disease. Nat. Commun. *9*, 5068.
21. Li, S., Zeng, W., Ni, X., Liu, Q., Li, W., Stackpole, M.L., Zhou, Y., Gower, A., Krysan, K., Ahuja, P., et al. (2023). Comprehensive tissue deconvolution of cell-free DNA by deep learning for disease diagnosis and monitoring. Proc. Natl. Acad. Sci. USA *120*, e2305236120.
22. Varley, K.E., Gertz, J., Bowling, K.M., Parker, S.L., Reddy, T.E., Pauli-Behn, F., Cross, M.K., Williams, B.A., Stamatoyannopoulos, J.A., Crawford, G.E., et al. (2013). Dynamic dna methylation across diverse human cell lines and tissues. Genome Res. *23*, 555–567.
23. Ziller, M.J., Gu, H., Müller, F., Donaghey, J., Tsai, L.T.Y., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D.A., Bernstein, B.E., et al. (2013). Charting a dynamic DNA methylation landscape of the human genome. Nature *500*, 477–481. https://doi.org/10.1038/nature12433.
24. Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. *30*, 207–210. https://doi.org/10.1093/nar/30.1.207.
25. Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L., et al. (2011). High density dna methylation array with single cpg site resolution. Genomics *98*, 288–295.
26. Mungall, C.J., Torniai, C., Gkoutos, G.V., Lewis, S.E., and Haendel, M.A. (2012). Uberon, an integrative multi-species anatomy ontology. Genome Biol. *13*, R5. https://doi.org/10.1186/gb-2012-13-1-r5.
27. Du, P., Zhang, X., Huang, C.C., Jafari, N., Kibbe, W.A., Hou, L., and Lin, S.M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinf. *11*, 587. https://doi.org/10.1186/1471-2105-11-587.
28. Chang, A., Jeske, L., Ulbrich, S., Hofmann, J., Koblitz, J., Schomburg, I., Neumann-Schaal, M., Jahn, D., and Schomburg, D. (2021). BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Res. *49*, D498–D508. https://doi.org/10.1093/nar/gkaa1025.
29. Yao, T., and Allen, G.I. (2021). Feature Selection for Huge Data via Minipatch Learning. Preprint at arXiv. http://arxiv.org/abs/2010.08529.
30. Doi, A., Park, I.H., Wen, B., Murakami, P., Aryee, M.J., Irizarry, R., Herb, B., Ladd-Acosta, C., Rho, J., Loewer, S., et al. (2009). Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. Nat. Genet. *41*, 1350–1353. https://doi.org/10.1038/ng.471.
31. Irizarry, R.A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., et al. (2009). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat. Genet. *41*, 178–186. https://doi.org/10.1038/ng.298.
32. Slieker, R.C., Bos, S.D., Goeman, J.J., Bovée, J.V., Talens, R.P., Van Der Breggen, R., Suchiman, H.E.D., Lameijer, E.W., Putter, H., Van Den Akker, E.B., et al. (2013). Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array. Epigenetics Chromatin *6*, 26. https://doi.org/10.1186/1756-8935-6-26.
33. Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70. https://doi.org/10.1038/nature11412.
34. Koestler, D.C., Jones, M.J., Usset, J., Christensen, B.C., Butler, R.A., Kobor, M.S., Wiencke, J.K., and Kelsey, K.T. (2016). Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). BMC Bioinf. *17*, 120. https://doi.org/10.1186/s12859-016-0943-7.
35. Moss, J., Magenheim, J., Neiman, D., Zemmour, H., Loyfer, N., Korach, A., Samet, Y., Maoz, M., Druid, H., Arner, P., et al. (2018). Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nat. Commun. *9*, 5068. https://doi.org/10.1038/s41467-018-07466-6.
36. Yang, J., Zhang, L., Yu, C., Yang, X.F., and Wang, H. (2014). Monocyte and macrophage differentiation: circulation inflammatory monocyte as biomarker for inflammatory diseases. Biomark. Res. *2*, 1. https://doi.org/10.1186/2050-7771-2-1.
37. Levine, M.E., Lu, A.T., Quach, A., Chen, B.H., Assimes, T.L., Bandinelli, S., Hou, L., Baccarelli, A.A., Stewart, J.D., Li, Y., et al. (2018). An epigenetic biomarker of aging for lifespan and healthspan. Aging *10*, 573–591. https://doi.org/10.18632/aging.101414.
38. Zhou, W., Laird, P.W., and Shen, H. (2017). Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. *45*, e22. https://doi.org/10.1093/nar/gkw967.
39. Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D., and Irizarry, R.A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics *30*, 1363–1369. https://doi.org/10.1093/bioinformatics/btu049.
40. Pidsley, R., Y Wong, C.C., Volta, M., Lunnon, K., Mill, J., and Schalkwyk, L.C. (2013). A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genom. *14*, 293. https://doi.org/10.1186/1471-2164-14-293.
41. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.
42. Fang, Z., Liu, X., and Peltz, G. (2023). GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. Bioinformatics *39*, btac757. https://doi.org/10.1093/bioinformatics/btac757.
43. Seabold, S., and Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In 9th Python in Science Conference.
44. Teschendorff, A.E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., and Beck, S. (2013). A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics *29*, 189–196. https://doi.org/10.1093/bioinformatics/bts680.
45. Chen, Y.a., Lemire, M., Choufani, S., Butcher, D.T., Grafodatskaya, D., Zanke, B.W., Gallinger, S., Hudson, T.J., and Weksberg, R. (2013). Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics *8*, 203–209. https://doi.org/10.4161/epi.23470.
46. Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. *34*, D590–D598. https://doi.org/10.1093/nar/gkj144.
47. Gamer, M., Lemon, J., Fellows, I., and Singh, P. (2019). irr: Various Coefficients of Interrater Reliability and Agreement. https://doi.org/10.32614/CRAN.package.irr.
48. Frankish, A., Diekhans, M., Ferreira, A.M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47*, D766–D773. https://doi.org/10.1093/nar/gky955.
49. Nixon, J., Dusenberry, M., Jerfel, G., Nguyen, T., Liu, J., Zhang, L., and Tran, D. (2020). Measuring Calibration in Deep Learning. Preprint at arXiv. http://arxiv.org/abs/1904.01685.
50. FOXO Bioscience . Methylize. . URL: https://life-epigenetics-methylprep.readthedocs-hosted.com/en/latest/index.html.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Human DNA methylation dataset	This work	Hugging Face: https://doi.org/10.57967/hf/7186
Human breast cancer dataset (TCGA-\BRCA)	The Cancer Genome Atlas Network[33]	https://portal.gdc.cancer.gov/projects/TCGA-BRCA
Manifest	Zhou et al.[38]	https://github.com/zhou-lab/InfiniumAnnotationV1/raw/main/Anno/HM450/HM450.hg38.manifest.gencode.v36.tsv.gz
Software and algorithms		
Python (v3.8)	Python Software Foundation	https://www.python.org/downloads/release/python-380/
R (v4.3)	R Core Team	https://cran.r-project.org/
NCBI eutils	NCBI	https://www.nlm.nih.gov/dataguide/eutilities/utilities.html
minfi (v1.44.0)	Aryee et al.[39]	https://bioconductor.org/packages/release/bioc/html/minfi.html
wateRmelon (v2.4.0)	Pidsley et al.[40]	https://bioconductor.org/packages/release/bioc/html/wateRmelon.html
minipatch learning (v0.1)	Yao and Allen[29]	https://github.com/DataSlingers/minipatch-learning
scikit-learn (v1.0.2)	Pedregosa et al.[41]	https://scikit-learn.org/stable/
gseapy (v1.1.3)	Fang et al.[42]	https://pypi.org/project/gseapy/
Methylize (v1.1.1)	Foxo Bioscience	https://pypi.org/project/methylize/
Statsmodels (v0.12.2)	Seabold and Perktold[43]	https://www.statsmodels.org/stable/index.html
irr (v0.84.1)	irr R package	https://cran.r-project.org/web/packages/irr/index.html
ontology-aware classification scripts	This work	https://github.com/ylaboratory/methylation-classification and https://zenodo.org/records/17861095

METHOD DETAILS

Sample downloading and curation

We used NCBI eutils to fetch and compile metadata for 59,123 human samples deposited on the Gene Expression Omnibus using the Illumina HumanMethylation 450K BeadChip (450K; platform: GPL13534) as of October 2024. Using the title and description fields of the metadata, for each sample we manually assigned a tissue or cell type label, a disease state (healthy or diseased), and treatment status (treated or untreated). Of those, we filtered samples using the following criteria: raw idat file availability, non-diseased status, and absence of experimental perturbation such as drug treatment. We further disambiguated tissue and cell type annotations by manually assigning them to the most descriptive tissue or cell term in the UBERON ontology[26] and merging functionally and physiologically similar terms, such as ‘‘buccal mucosa’’ and ‘‘oral epithelium’’ (Table S8).

DNA methylation data preprocessing and normalization

Raw Illumina 450K array data were processed into beta values and background corrected using the standard Noob (normal-exponential out-of-band) method implemented in the minfi package.[39] Preliminary sample quality control excluded 81 samples based on median intensity values across control probes, reducing the final set of samples to 16,959. To account for the differences between type 1 and type 2 probes in 450K data, beta values were normalized using beta mixture quantile dilation (BMIQ) normalization[44] from the wateRmelon package.[40] Probe-level quality control was performed using detection p-values (cutoff = 0.01),[39] and probes associated with single nucleotide polymorphisms,[39] cross-reactive probes,[45] and those located on the sex chromosomes[39] were removed, narrowing down the total number of probes to 297,598. Finally, beta values were converted to M-values, then used for downstream classification tasks. For the independent GEO: GSE77797[34] evaluation dataset with blood samples of known cell

composition, rather than processing from raw array data, we directly converted provided beta values to M-values. Infinium MethylationEPIC (850 K) samples from GEO: GSE122126[35] and additional diseased and adjacent normal 450K samples from TCGA-BRCA[33] for generalizability analyses were preprocessed using the same process. Any probe values measured in 450K but missing from 850K were imputed with median sample methylation values.

Sample partitioning for training or label transfer validation

We then partitioned all remaining annotated samples into either the training set or the label transfer validation set. To ensure robust coverage in the training set, we required each tissue or cell type label to be present in at least 2 independent studies, with a minimum of 2 samples per study, and a minimum of 6 total samples per label. Then, to ensure comprehensive coverage of physiological systems in the ontology, we selected a subset of children labels not passing the training set criteria to augment system nodes with very low training sample set sizes, including ''cardiovascular system'', ''sensory system'', ''excretory system'', and ''nervous system'', resulting in a set of 10,351 samples (Table S1). The remaining tissue labels that did not meet the coverage criteria were used in the label transfer validation set. Samples annotated generally as ''blood'' without cell types were also used as part of the label transfer validation set.

Genomic annotation of CpG probes

To link methylation probes with genomic and transcriptional features, we mapped Illumina hg19 coordinates to hg38 using LiftOver.[46] Probes were annotated to CpG island, shelf, shore, and open sea as defined by the Illumina manifest: islands as regions with length >500 bp and >55% GC, shores as <2kb from islands, shelves as <2kb from shores, and the remaining as open sea.[38]

Tissue and cell ontology

In order to systematically define the physiological relationships between tissue and cell type labels, we leveraged the extended UBERON ontology,[26] which includes the Cell Ontology, together with the BRENDA tissue ontology,[28] which captures functional anatomical tissue relationships. We manually curated each ontology edge for all tissue and cell type labels in our complete data compendium (including both training and label transfer sets) as well as indirectly connected entities either through neighboring or intermediate connections between observed sample annotations. Specifically, the edge curation was based on existing *is_a*, *part_of*, and *develops_from* relationships in the UBERON and BRENDA ontologies to build a biologically intuitive directed acyclic graph (DAG) primarily organized by organ system. This resulted in a DAG with 118 nodes and 139 edges including the root node. To obtain the smaller training ontology, we subsetted entries of the ontology to contain only entities for tissues or cell types in the training set or associated indirect nodes, resulting in a training DAG with 72 entities joined by a root node and 88 edges.

CpG feature selection using minipatch learning

For feature selection, we used the minipatch learning method.[29] Briefly, minipatch learning iteratively selects random subsets of features and samples, referred to as 'minipatches' and assesses feature importance using decision trees for multiclass tissue and cell type classification. The resulting feature importance from each patch are then combined with cumulative importance values, which then inform feature selection probabilities for subsequent minipatches. Through this iterative sampling approach, each feature's utility across the entire dataset can be efficiently estimated based on its overall selection frequency, enabling identification of a small set of highly informative CpGs.

Runtime evaluations

Runtime evaluations for both differential methylation and minipatch learning for feature selection were restricted to 10 threads on a server with 4 Intel(R) Xeon(R) Gold 5220 CPUs and 1.5TB RAM. Only library-internal parallelization was permitted, thus differential methylation was performed using parallelization and minipatch learning was not. For evaluation purposes, we created subsampled datasets with 100, 500, 1,000, 5,000, and 10,000 samples from the entire training set, stratifying by tissue labels to maintain consistent representation of tissue labels across runtime measurements. Measured runtimes include only the time required to fit each feature selection method and not downstream classifier training and prediction.

QUANTIFICATION AND STATISTICAL ANALYSIS

Within tissue or cell type sample consistency evaluation

To measure the variation among directly annotated samples from the same tissue or cell type in the training set, we calculated the intraclass correlation coefficient (ICC), a statistical measure that evaluates the consistency of observations within defined groups. ICC values were calculated using the two-way consistency model based on single-measurement units implemented in the irr R package.[47] Higher ICC values indicate greater consistency among samples, reflecting stable tissue- or cell-type-specific methylation signatures.

Hyperparameter optimization and cross-validation

Hyperparameters for feature selection and classification were set to default values, with the exceptions of minipatch size and selection frequency threshold. Recommended minipatch learning parameters for the size of sample and feature ratios per minipatch were used ($\frac{\sqrt{(N)}}{N}$, where N is the total number of samples or features, respectively).[29] Selection frequency threshold was optimized using 3-fold cross-validation implemented with scikit-learn.[41] Cross-validation folds were stratified by tissue labels and grouped by dataset, ensuring that samples from a single dataset were restricted to the same fold to avoid data leakage. Within each fold, we assessed minipatch learning and classification performance using F1 score as the evaluation metric. We determined the selection frequency threshold by identifying the elbow point of classification performance across cross-validation folds (selection frequency = 0.65), and a final minipatch learning and multi-label classifier with optimized hyperparameters on the entire training dataset were used for downstream predictions.

Gene enrichment of selected probes

To characterize the biological functions associated with the CpGs selected by our feature selection framework, we mapped each probe to its associated gene(s) using the known probe to GENCODE[48] associations from the microarray manifest.[38] Unique genes corresponding to the selected probes were then subjected to hypergeometric enrichment analysis using gseapy[42] across multiple libraries, including GO Biological Process, Molecular Function, and Cellular Component 2023, KEGG 2021, and MSigDB Hallmark 2020. The set of all genes represented in the training data associated with probes that passed quality control for the 450K array platform was used as the background for enrichment, ensuring that significance was assessed relative to the assay's gene coverage. Enrichment results of pathways with adjusted p value less than 0.05 were considered as significant and were reported.

Ontology-aware multi-label classification

To effectively leverage hierarchical relationships among tissues and cell types present in our ontology, we propagated the node labels through the DAG, such that each sample's annotation includes not only its directly annotated label, but also any parent labels. Thus, our classifier could learn methylation patterns associated with both precise tissue and cell types as well as more general signals, including for organ system level nodes. For classification, we used a multi-label support vector machine (SVM) with balanced class weights and a linear kernel, implemented via scikit-learn.[41] Predicted probabilities were calibrated using Platt scaling[41] to reflect the empirical fraction of positive samples. Labels with predicted probabilities above 0.5 were considered positive, while cases where no label exceeded this threshold were reported as 'no prediction.' The calibration quality was evaluated using expected calibration error (ECE),[49] confirming that the predicted probabilities are well-calibrated and suitable for interpretation.

We define n as the number of samples, indexed by i, with true and predicted label sets $\mathbf{y}_i$ and $\widehat{\mathbf{y}}_i$, respectively. Sample-wise precision (Equation 1) averaged per-sample scores, while tissue-wise precision (Equation 2) and F1-score (Equation 3) were computed per tissue and summarized by the median. Tissue-wise metrics were only computed for tissues with at least one positive sample.

$$\text{Precision}_{\text{sample}-\text{wise}} = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FP_i} \quad \text{(Equation 1)}$$

$$\text{Precision}_{\text{tissue}-\text{wise}} = \text{median}\left(\frac{TP_t}{TP_t + FP_t}\right), t \in \{1, \ldots, T\}, \text{where } TP_t + FP_t > 0 \quad \text{(Equation 2)}$$

$$\text{F1}_{\text{tissue}-\text{wise}} = \text{median}\left(\frac{2TP_t}{2TP_t + FP_t + FN_t}\right), t \in \{1, \ldots, T\}, \text{where } TP_t + FN_t > 0 \quad \text{(Equation 3)}$$

Differential methylation baseline

Differential methylation was used as a baseline feature selection method due to its widespread use as a standard analytical approach in DNA methylation studies, particularly for identifying tissue-specific methylation markers. We calculated tissue-specific differentially methylated probes via one-versus-rest comparisons while accounting for dataset labels as covariates using the Python package Methylize.[50] Probes were considered significantly differentially methylated if they passed an alpha of 0.05 after multiple hypothesis test correction using the Statsmodels package.[43]

For machine learning-free, correlation-based classification based on differential methylation, we used the union set of differentially methylated probes across all tissues. Within the established cross-validation folds, we calculated sample-to-sample Pearson correlation coefficients between samples in training folds with those in the validation fold. Tissue and cell type predictions for each sample in the validation fold were assigned based on the tissue or cell type label corresponding to the highest average correlation coefficient of grouped training samples. The metrics reported are averages across all cross-validation folds.

Ablation studies

Ablation analyses were performed under three configurations: (1) ablation of feature selection, (2) ablation of classification framework, and (3) ablation of both. For feature-selection ablation, we generated a comparison probe set based on differentially

methylation. For each tissue, probes were ranked by its significance (1-vs-all differential methylation), and the top-ranked probes were assembled until matching the size of the minipatch learning-selected input probe set. This differentially methylated (DM) probe set was then used to train and test the ontology-aware multi-label framework using the same data splits as the main analysis. For the classification-framework ablation, our minipatch learning-selected probes were used to train and test conventional multiclass machine learning classifiers: Elastic Net, random forest, and multilayer perceptron (MLP).[41] When both feature selection and classification framework were ablated, the previously mentioned top differentially methylated probe set was used to train and test the same multiclass classifiers.

Label transfer evaluation

To assess prediction performance in the label transfer evaluation, we devised a graph distance-based metric for each sample that measures the average distances of each predicted label generated by the multi-label classifier to the annotated true label. Specifically, we used an undirected version of our ontology that included both the training and unseen labels (Figure S2). Because predictions would be assigned to labels present in the training set, we computed an adjusted ontology distance that took into account the nearest available training node for each unseen label. More formally, let d_i represent the ontology distance between the target (unseen) label and the i-th predicted label for a given sample. We first identified the minimum distance between the target node and all n nodes in the training set:

$$d_{\text{min}} = \min(d_1, d_2, \ldots, d_n).$$

We then computed the adjusted ontology distance by subtracting this minimum achievable distance from each prediction's ontology distance:

$$d_{i,\text{adj}} = d_i - d_{\text{min}}.$$

Thus, an adjusted distance of 0 indicates an optimal prediction (matching the closest possible training node), while adjusted distances reflect less accurate label transfer.

To contextualize our label transfer performance, we established random baseline distributions for each unseen label. This involved randomly sampling labels with replacement from the full set of labels in the training ontology (1,000 iterations per unseen label). For each sampled label, we calculated the adjusted ontology distance to the target unseen label using the same method described above.